Integrating Data from Natural Language Processing into a Clinical Information System

Stephen B. Johnson, Ph.D [1]
Carol Friedman, Ph.D. [1,2]

[1] Department of Medical Informatics
Columbia University, New York

[1,2] Department of Computer Science
Queens College of CUNY, New York

Demographic data extracted from discharge summaries by natural language processing was compared to data gathered by a conventional hospital admitting system. Discrepancies in data were noted in names, age, sex, race, and ethnicity. Some differences are attributable to errors in collection: interaction with patient, dictation, transcription, and data entry. Very few differences were due to errors in natural language processing. Other differences can be used to critique existing data, or to enhance data with more detailed information. Discrepancies in data as elementary as patient demographics raise the issue of resolving conflicts when neither source of data is known to be more reliable. Clinical repositories can represent conflicting data from multiple sources, but clinical information systems must bear the cost of increased complexity in the application programs that will use the data.

INTRODUCTION

The Clinical Information System (CIS) at Columbia-Presbyterian Medical Center (CPMC) collects data from multiple clinical applications into a central clinical repository [1, 2]. The repository currently holds data for 1.3 million patients covering a period of roughly five years. The bulk of the data is laboratory results, medication orders, radiology procedures, demographic data, and text reports (radiology, pathology, cardiology, discharge summaries, etc.).

Recently, CPMC has begun using natural language processing (NLP) techniques to analyze free text reports, extract relevant clinical information, and store it in the central repository. Radiology reports were the first to be analyzed in this way, and are now being routinely processed in the production system [3]. Evaluation of this system has demonstrated that NLP is a reliable technology for use in a clinical information system in conjunction with decision support applications [4].

Attention is now be directed toward the processing of discharge summaries. Summaries are complex documents compared with the reports produced by diagnostic procedures such as radiology. One aspect of this complexity is that discharge summaries constitute a particular **view** of clinical data, in the sense that they are synthesized from multiple **base** events that are much simpler in structure. In many cases, these base events are also captured by other computer applications in the institution (laboratory data, medications, etc.).

NLP decomposes complex documents into basic events that can be stored in the clinical repository. While the data made available by this processing will certainly be extremely valuable, problems in the consistency of data may arise. An example of this situation is shown in Figure 1. When a patient is admitted to a hospital, a history is taken by a physician, and the patient is also interviewed by staff as part of the admission process. Information collected by physician is ultimately dictated in the form of a discharge summary, which is transcribed, processed by the medical records system, and sent to the central clinical repository for storage. Admitting personnel collect demographic and other data from the patient and enter it into the admitting system, which sends it to the repository. For both textual and structured data, the Health Level 7 (HL7) standard is employed when data is uploaded to the central repository [5].

The new component in this process is natural language processing, which extracts demographic and other information from the text for storage into the clinical repository. Thus, there are now two sources for demographic information, which introduces the possibility of inconsistency. Since the purpose of the clinical repository is to provide an integrated view of a patient's data, it is important to understand in what ways data extracted by natural language processing differ from data collected in conventional ways. If NLP can be shown to be accurate, a method of reconciling these differences must be established.

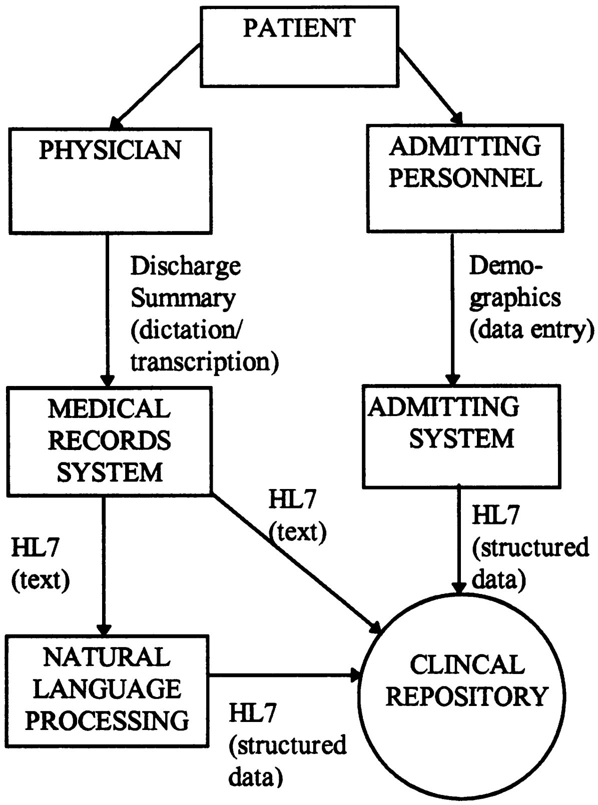

Figure 1

Multiple sources of demographic information.

In a heterogeneous information system like the one at CPMC, data are often initially collected to serve a local need. When integrated into the clinical repository, these data are made available for use by many other applications with very different purposes. For example, the demographic information collected by an admitting system serve to identify a patient. These data may also be useful for patient care. Attributes such as race or ethnicity may have only marginal value for patient management or care, but may be extremely interesting for research purposes, such as correlation of cancer information with genetic and social factors.

To begin to address the issues of data accuracy and level of detail, a simple experiment was conducted to compare demographic data extracted through natural language processing against data gathered by the CPMC admitting system. This experiment is a first attempt at answering the following questions:

- Can NLP be used to analyze discharge summaries in a production clinical information system?
- Can the data extracted by NLP be used to critique the quality of data gathered by conventional computer applications?
- Can NLP data be used to enrich patient data for use by clinical research applications?

METHODS

640 discharge summaries from 1995 were selected at random. The summaries were processed by the MedLEE system [3]. Only paragraphs labeled as "history of present illness", "summary", "medical summary", and "chief complaint" were analyzed. The NLP system was originally developed to process radiology reports, but only underwent minimal training before being applied to discharge summaries. A new semantic class for "ethnicity" was added to the grammar, and lexical items for ethnicity were added to the lexicon. The rule for processing numbers had to be extended to handle numbers in age expressions that are written out in full, such as "sixty seven years old". Extracted data was represented in list structures, as depicted in Figure 2.

MRN: 1234567
NAME: ROSS, BETSY
DISCHARGE DATE: 04-17-95

[finding, demo, [age, [unitval, 48, year]], [race, white], [sex, female]]
[finding, asthma, [status history]]
[finding, coronary artery disease]
[finding, hypertension]
[finding, cigarette smoker]
[finding, positive family history]
[finding, myocardial infarct, [status, previous]]
[finding, diabetes, [certainty, no]]
[finding, hypercholesterolemia, [certainty, no]]

Figure 2

Findings extracted by natural language processing.

A program written in the Perl computer language was written to collect the header information (medical record number, patient name, discharge date), and demographic findings in the extracted data, and organize them into a tabular format to facilitate further comparison. The extracted data were compared (by manual review) with the original discharge summary text to assess the accuracy of the natural language processing. A given datum, e.g. race, could be missing the in text (not reported), reported and correctly extracted by MedLEE, or reported but not correctly extracted.

The data judged to be correctly extracted were then compared with demographic data stored in the central clinical repository (a relational database implemented in DB2 on an IBM mainframe). The medical record number (MRN) was used to construct a query in Structured Query Language (SQL) to access the PATIENT table in the database. The following data elements were compared: MRN, last name, first name, age, sex, race, and ethnicity. The patient age given in the summaries was adjusted by accounting for the date of discharge.

RESULTS

Manual review of the discharge summaries showed that the MedLEE system had no difficulty extracting MRN, patient name, age, race, or ethnicity when these elements were present in the paragraphs analyzed. (They were sometimes present in other paragraphs not analyzed as indicated below.) Some difficulties occurred in determining the sex of the patient, as discussed below. Thus, for demographic data other than sex, accuracy of the MedLEE system was not an issue for this study.

When data extracted by MedLEE were compared with data in the central repository, a demographic attribute could fall into one of four different categories: it could be missing in the discharge summary, it could be missing in the admitting system data, the value could match in both of these systems, or the value could fail to match. Table 1 summarizes missing and mismatched data elements. Mismatches will be reported below as a percentage of the total number of comparable records (i.e., data that are present in both data sets).

One medical record number (0.1%) was found to be missing from the clinical repository, meaning that the central system had no entry for a patient with that identifier. 37 last names (5.7%) in the headers of the free text reports did not match the name in the repository for the given MRN. Name mismatches were further analyzed as follows. 16 names (2.5%) were found to be alternate names for the patient by consulting the PATIENT_ALIAS table in the repository. 20 names (3.1%) were found to match within a specified tolerance, as defined by the "least common substring" (LCS) algorithm, which was used to compute similarity of patient names [6]. One name (0.1%) was determined to fail LCS matching (the significance of this failure is discussed below).

Of the remaining records, in which the last name matched exactly, 39 first names (6.0%) from the headers of the summaries did not match the name in the repository. All of these were judged similar by LCS matching, except one (0.1%).

Large differences were found in patient ages. The number of discharge summaries containing differences of more than 1, 5, 10, or 50 years are listed in Table 2.

Age Diff	# Records	%
1	72	14.8
5	19	3.9
10	8	1.6
20	6	1.2
50	3	0.6

Table 2

There were four records (0.8%) in which the sex of the patient did not match. In two cases, the NLP system seems to have made a mistake. In one of these, the NLP

Category	MRN	LAST	FIRST	AGE	SEX	RACE	ETHNICITY
Missing in Discharge Summary	0	0	1	152	130	564	578
Missing in Admitting Data	1	0	0	0	0	11	640
Present in both data sets	639	640	639	488	510	553	0
Mismatch	0	37	39	72	4	2	0

Table 1

Comparison of demographics in discharge summaries and admitting data.

system mistook the sex of a pregnancy test for the sex of the patient. In the remaining two cases, the admitting data seems to be in error, by identifying the sex of the patient as female, while the text of the summary uses words such as "he" and "man".

In 2 records (0.4%) there was a difference in the race reported for the patient. In the first, the NLP system identified the patient as white, while the repository used the race code "other". In the second, the NLP system identified the patient as black, while the repository used the code for white. There were 11 records (1.7%) in which race in the clinical repository was unknown, while NLP was able to determine race for the patient. In 62 records (9.6%), the NLP system was able to determine an ethnic group for the patient.

DISCUSSION

In the highly heterogeneous environment at CPMC, in which computer applications are acquired from multiple vendors and run on a wide variety of platforms, complete integration of patient information is still far from perfect. Medical record numbers are occasionally entered incorrectly, causing failures when data are uploaded from an application system to the clinical repository. These failures are reported back to the application for correction. Because a medical record number may match spuriously, it is also necessary to match on the patient's name for confirmation.

In some applications, patient names are treated like an additional data item and are re-entered. Because re-keying of patient names may result in misspellings, an important part of the upload architecture at CPMC is tolerance of name variation within specified limits. This is achieved by use of the "least common substring" (LCS) algorithm, which was found to be highly successful in accepting names that would be judged similar by humans [6]. When LCS fails to match a patient name against the central system, this taken as strong evidence that the MRN is erroneous. In these cases, the information being uploaded is rejected, and the error reported to the sending application.

This architecture entails that the majority of discrepancies in patient names between the admitting system and discharge summaries would have no real effect on the upload process, and each discharge summary would be associated with the correct patient record. A mismatch in medical record number, or severe mismatch in name will cause the discharge summary to be returned to the medical records system for correction.

At present, the algorithm for determining whether two patients are the same person or not does not take age or sex into account. The results indicate that even though discrepancies in sex are fairly low, mistakes do occur. Examination of age data in discharge summaries indicates that there is a certain degree of natural variation. This may be partly due to the fact that the patient's age is collected, rather than the exact date of birth, as is done in the admitting system. Some large differences are clearly the result of typographical error (e.g., one patient's age was entered as 655 years). However, other large age differences remain which are not easily explained by errors in data collection. For example, in one case an age difference of more than 50 years is accompanied by a difference in sex as well. This suggests that checking age may also be useful in determining patient identity, and that very large discrepancies be considered suggestive of a patient mismatch.

In cases in which race is unknown in the repository, information from the discharge summary may be suitable in its place. While the percentage of cases in which this was possible was low (1.7%), analysis of more paragraphs of the summary (e.g., "social history") are likely to yield additional information. Ethnicity is not well represented in the repository (it was treated as missing in this experiment). The experiment showed that 9.6% of the records could enhance ethnic information about patients. It is reasonable to expect that this can be increased by another 10% by analyzing the "social history" paragraph in the discharge summary, which often indicates country of origin, language spoken, etc.

While enhancements in race and ethnicity may hold little interest for patient administration and clinical purposes, researchers at CPMC have often complained about the poor quality of race and ethnic data in the repository, in particular with regard to the lack of detail in ethnicity (e.g., recording Hispanic rather than Dominican). The addition of more detailed information may aid in establishing correlations of disease data with genetic and social factors.

The wealth of information in discharge summaries suggests great possibilities for future applications. The data in Figure 2 provides an example of the kinds of findings that can currently be obtained through natural language processing. Further training on discharge summaries will increase the spectrum of data that can be reliably extracted.

However, the above comparison of data indicates that there will be discrepancies even in data as elementary as demographic attributes. This suggests that differences in more complex clinical observations will be even greater. For some data, available coded data (e.g., ethnicity) may be considered virtually worthless, and the addition of any new values will be welcomed. For other data (e.g., age), secondary sources may be too unreliable to be used to alter existing data. Secondary sources of data may aid in performing quality assurance of primary data, e.g., when age differences are extreme.

The most difficult obstacle to integrating clinical data occurs when there are two or more sources of a datum, and no one source can be considered more reliable than the others. The clinical repository can model this situation using a structure such as Table 3, in which the same datum can be reported by different sources.

MRN	TIME	SOURCE	DATUM	VALUE

Table 3

Schema for multi-source clinical observations.

This type of schema is consistent with the current information systems architecture philosophy at CPMC to make as much clinical information available as possible from as many sources as possible. While this design facilitates rapid incorporation of new sources of information into the clinical information system, it places a tremendous burden on applications that must retrieve data from the repository: each application must make judgments about which sources to trust when there is a conflict in the values of a datum for a given time period.

As long as natural language processing applications are still new and relatively untested, storing extracted data in parallel with data collected by existing applications allows the information to be used immediately by those who wish to put faith in it, or not used by those who still have doubts. As formal evaluations demonstrate that NLP applications can be considered robust and reliable, the programs that manage the clinical repository must be made much more intelligent; differences in data will have to be weighed according to well-defined rules to produce an integrated patient record for use by other applications.

References

1. Johnson SB, Forman B, Cimino JJ, Hripcsak G, Sengupta S, Sideli R, Clayton PD. A Technology Perspective on the Computer-base patient record. In: Steen Ed, ed. First Annual Nicholas E Davies CPR Recognition Symposium. Washington DC: Computer-based Patient Record Institute, 1995:35 -51.

2. Johnson SB, Hripcsak G, Chen J, Clayton PD. Accessing the Columbia Clinical Repository. In: Ozbolt J, ed. Eighteenth Symposium on Computure Applications in Medical Care. Philadelphia: Hanley and Belfus, 1994:281-5.

3. Friedman C, Alderson PO, Austin HM, Cimino JJ, Johnson SB. A General Natural Language Text Processor for Clinical Radiology. JAMIA, 1994;1(2).

4. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing. Annals of Internal Medicine 1995;122(9):681-8.

5. Sideli R, Johnson S, Weschler M, Clark A, Chen J, Simpson R, Chen C. Adopting HL7 as a standard for the exchange of clinical text reports. In: Miller R, editor. Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care; 1990 Nov 4-7; Washington, D.C.; 1990: 226-229.

6. Friedman C, Sideli RV. Tolerating spelling errors during patient validation. Computers and Biomedical Research. 1992;25;486-509.